# WOODHOUSE EXHIBIT 3

# EXHIBIT D

# feel Data update for GenAI

Document status: collecting content | draft | **final**

**Share with:** GenAI leadership

**Invited by:** Jort Gemmeke

**Participants (DI):**

| | |
|---|---|
| Delia David | TL DI/Data4AI |
| David Levin | PM DI/Data4AI/Training Data |
| Barak Yagour | Director DI/Data4AI |
| Jelena Pjesivac-Grbovic | Director DI/Data4AI/Training Data |
| Maor Kleider | PM Director DI |

## Meeting Objectives

1. **Inform** on the problems we identified and the plan to address them
2. **Present** new product demo: AI Data Catalog (beta)
3. **Discuss** on future opportunities and priorities

## Scope

The scope of this document is to discuss some of the Data related problems & initiatives under the broader X-Infra workstreams kicked off in April-23'. With specific drill down into topics related to Data Cataloging, Data Preparation and Compute Engines.

| X-INFRA GenAI Workstreams | | | | |
|---|---|---|---|---|
| **WS1 - SUPPORT CONNECT LAUNCH** | Trainers & Scheduling | Inference | Capacity | |

| WS2 - LONG-TERM INVESTMENTS TO SUPPORT GENAI | Research Cluster in Infra: Software | Research Clusters in Infra: HW | Capacity Planning & Handoff | Training & Scheduling | Data & Storage | Authoring/ Researcher DevEx | Monitoring & Observability |
|---|---|---|---|---|---|---|---|

Considering the emergent, nascent and rapidly evolving landscape of Generative AI, we are adopting a fast-iterative approach. Infra will continue to closely partner with the GenAI team, ensuring rapid adaptability. While joint strategic investments will remain a priority, we anticipate changes to our execution plan based on the evolving GenAI's needs.
Below are the H2 focus areas as of August 2023.

## 1. Optimizing training data loading

**Goal: Optimize data loading performance and ensure it is not a bottleneck for GenAI training. Enhance flexibility of data loading strategies.**

Earlier in H1, when several Generative AI and Content Understanding workloads across GenAI, RL, Ads started reading from Everstore and Manifold, we experienced slow data reads, throttling and reliability problems. This is due to a combination of capacity allocation, lack of data and compute collocation (within the same datacenter campus), and lack of data layout optimization.

AI Infra and Data Infra have greatly reduced the issues above for Ranking and Recommendations use cases through data & compute (GPU) colocation for thousands of training datasets, and have improved efficiency and training QPS with investments in data layouts and optimal data loading during training. We need to do the same for GenAI workloads to ensure data loading does not slow down or fail GenAI training. This requires us to understand dataset composition and access patterns (single and multimodality, sparse rows, partial vs. entire data access such as few frames vs. full video, sequential vs. random reads), and the type and amount of compute required (across a heterogeneous GPU fleet), in order to automatically decide where to place data and enable high-throughput reads during training.

We have already copied a large amount of datasets (especially media ones) to the data warehouse, which enabled the colocation with GPUs, and we will continue to invest in the following:

1. Enhance the data placement algorithms to support composite (multiple tables) and multimodal GenAI datasets.
2. Continue to evolve DPS (Data Preparation Service) for LLM & media data ingestion and media data transformations (i.e. media resizing, video frame sampling, video & audio alignment).
3. Improve data loading reliability and debuggability. Support various data loading strategies, i.e. deterministic, shuffled, sampled.

**Commented [1]:** can i get some steer and can we touch on how our media and language folks on how this has helped us so far?

**Commented [2]:** this reads like it is not done and we need to?

**Commented [3]:** where are we on language?

## 2. Enable dataset management and discovery

**Goal: Unified data management and discovery experience for GenAI datasets, with automatic metadata collection and robust privacy compliance.**

We are evolving the existing data management platforms, already proven at scale. The Hive Metastore is our solution for managing metadata for tens of millions of tables. Customers define dataset schemas, partitions, retention policies, and we have lineage systems, such as the Unified Lineage System and AI Metadata, integrated with all our execution engines, such as Spark and Presto, to capture how datasets, columns and partitions are used, and to enable explainability, legal enforcement, privacy decisions and enforcement.

Generative AI datasets are often composite (multiple tables built from multiple sources of data) and multimodal. Generative AI training candidates undergo mitigations (such as legal, copyright, safety, privacy, jurisdiction, integrity ones), metadata augmentations (such as language identification, entity detection, media aesthetics, media auto-captioning) and data transformations (such as masking, synthetic data generation, color scheme normalization, cropping, segmentation). Some of the same mitigations have already been used at small scale, for certain integrity and privacy purposes, however we did not have requirements in the past to add structure and discoverability for all the above. For Generative AI use cases, we have been asking ourselves what metadata should we collect for easy multi-table and multimodality dataset management and discovery, how to track data mitigations, transformations, augmentations and filters applied during data preparation, and how to ensure privacy compliance (such as data deletion and data usage policies).

Over the past months we have partnered with the GenAI team, especially the GAID (GenAI Data) team, to build the **AI Data Catalog.** This is also in partnership with RL and AI Infra, leveraging existing DAMIT (Data and Model Insights) investments, and making metadata and lineage available in AIM (AI Metadata). We have been focused on the following:

1. Register composite multimodal datasets and discover them by modality, sources of data, metadata augmentations, creation time, usage.
2. Manage dataset versions, integrate with training workflows, automatically track lineage, usage and privacy compliance.
3. Automatically collect data preparation, curation and filtering metadata for infra-supported data operators. Enable easy registration of such metadata from custom code.

**Commented [4]:** i think we should really optimize for:
1. scale
2. access on TC
3. streamlining pxfn decisions because of the data catalog effort.

**Commented [5]:** Do we have more details on this in terms of data flow and how does it fit with existing workflows?

**Commented [6]:** This is a new system more tuned for AI datasets vs the existing Metastore. Functionality built here will be around multi-table datasets, versioned access and metadata support for modality tracking / mitigations / annotations.

Given the context above, this system will be integrated in the data preparation and training workflows in H2. Please find a few more below (also please note that we review all these in a sprint based fashion with GenAI - and we have not reached this milestone yet, so please take this as tentative integration):

- Data Preparation: as data is ingested into our data warehouse, mitigated and transformed, metadata on these operations is automatically tracked when using Infra-provided Dataswarm Operators (which we will design and build together). For non-standard Operators, we will have APIs to register the same pieces of metadata.

- Training workflows: imagine using a dataset backed by multiple tables. Say we call it X. Training job owners would go to Mast and specify dataset X in their job config. Mast would resolve X into underlying tables by calling this new service. This way, customers do not need to remember and type 4 table names when using a dataset with 4 tables.

- Debugging in Presto / Spark: we have not planned this yet, but we might want to add an integration so that one can go to DataQuery and use a dataset name for querying vs the 4 underlying tables in the example [1]

**Commented [7]:** Curious if and how this interfaces with idata?

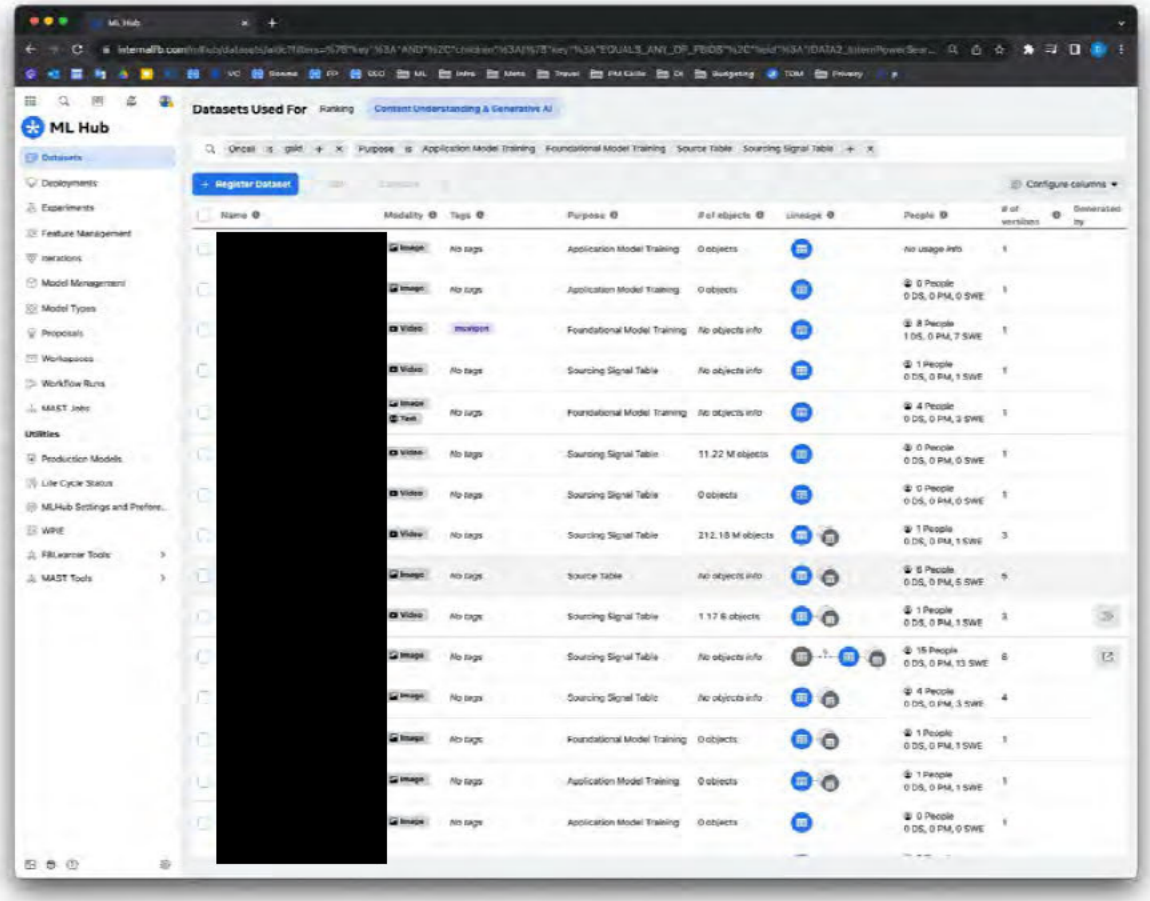
Commented [8]: Redacted - Privilege

Redacted - Privilege

1 total reaction [2]

**Commented [9]:** @[redacted]@meta.com - yes, this is possible and where we'd like to go. What would be useful is to understand the process in more details, worth a followup.

re iData - yes, this interfaces w iData and is actually built by few members of the iData team.


ML Hub
Datasets Used For
Register Dataset

## 3. Infra data preparation parity with AWS

**Goal: Enable data ingestion and preparation pipelines on prod Infra, through Open Source Spark, reusable Dataswarm Operators and dev experience parity with AWS.**

Data Warehouse has been dominated by internal structured/schematized data for the past 10+ years, sourced from user activity logs (Scribe) and the user database (UDB). SQL has been the language of choice for Data Warehouse customers and Dataswarm is the data pipeline authoring framework used for authoring millions of data tasks across the company. Dataswarm offers a set of popular Operators for moving data between storage systems, processing data, performing privacy checks, etc. Users of Dataswarm can author their own Operators, as well as custom python code and libraries, to express the data operations they need to perform. We use internal versions of Spark and Presto for executing Dataswarm Operators, optimized for the data processing needs we've had over the years. Spark, in particular, has forked from Open Source 4 years ago, and we have yet to migrate to the latest ML libraries and PySpark.

With the recent focus on Generative AI, which can benefit from using Facebook / Instagram / WhatsApp data combined with data sourced from the Internet and acquired data, we have been asking ourselves what new data ingestion mechanisms we need to build? What new Dataswarm Operators should we offer for generative AI data preparation? What compute engine capabilities are we missing?

The immediate requirements from GenAI are available in this doc from Jacob. We have recently refocused the team on **PySpark, Open Source,** and are in the process of defining what **Generative AI Dataswarm Operators** we should build. We are investing in:

1. Data transfer and ingestion operators, i.e. transfer data between AWS and Meta's Data Warehouse, ingest data from the Internet and ingest acquired data. Offer options for both ad-hoc and periodic/continuous data transfers.
2. PySpark support across Dataswam, Daiquery, enabling fast code migration from AWS to Infra. Migrate several pipelines, including CommonCrawl, in H2.
3. Popular GenAI data Operators, such as deduplication and UII removal / redaction.

## 4. Build data understanding insights and tools

**Goal: Enable researchers to make faster decisions about their datasets through insights into the data and data visualization and comparison.**

Data Infrastructure has always invested in the performance and efficiency of our growing use cases, and over the last few years the growth has been driven by ML use cases (75% of the warehouse data is ML data, and some of the datasets are hundreds of petabytes). Data quality and privacy / compliance investments have enabled users to debug data issues and make data decisions faster.

We will continue our investment in data understanding, with a focus on the new Generative AI use cases. We will be focusing on the following areas:

1. Allow researchers to identify shortcomings in their datasets (i.e. the lack of high-resolution mountain photos) and subsequently guide decisions on enhancing training and context data.

2. Out-of-the-box data quality insights, such as value distribution, skews and outliers. If there is a need, we can also enable users to specify or select from a list of algorithms for computing entity distribution across a taxonomy, safety, bias and responsible AI data insights.
3. Tools and products for visualizing and comparing datasets and versions, enabling a unified way to view data and metadata across all modalities.

## H2 Roadmap

| | Outcomes | Enabling Technology |
|---|---|---|
| **Today** | • Media datasets **colocation with GPUs**<br>• Ability to **offload media transformations** before training (i.e. resize, central crop, video frame extraction)<br>• **Media** datasets registered in Data **Catalog**<br>• **Data Lineage**, as captured by DI systems, available in Data Catalog | • **Tetris** and **DPS** (media - compute colocation and media transformations)<br>● **AI Data Catalog** |
| **Q3-23** | • Repeatable data transfers from **S3 to Infra**<br>• **First LLM pipeline**(s) migrated from S3 to Infra<br>• **PySpark beta** development available for testing<br>• **LLM** datasets registered in Data **Catalog**, including AI Agents Data Flywheel<br>• Data **Catalog** in the critical path of **training workflows (removing configs)**<br>• **Photo** dataset **visualization** (single view of photos and associated metadata) | • **Dataswarm** Operator for transferring S3 data to Infra, implemented in DPS<br>• **OSS Spark, PySpark** and ML libraries beta deployment<br>• **AI Data Catalog** integration with **MAST** and Data Loader(s) used by GenAI<br>● AI Data Catalog extension for **image** dataset **visualization** |

**Commented [10]:** How will this be different from DAMIT?

Furthmore, shouldn't DE/DS own this?

**Commented [11]:** Part of the visual capabilities and insights exist in DAMIT. DAMIT is working with AI Infra and Data Infra on building together and consolidating onto the new Data Catalog solution, which will offer higher scale and more capabilities over time.

I assume that a lot of data insights are being built by DE/DS today. We want to have a discussion on whether we see value in computing some of these by default for certain classes of datasets. This would only make sense if the same insights apply across a large number of datasets.

**Commented [12]:** We also look to expand the existing Media data understanding capabilities in DAMIT to other modalities including multimodal and LLM. We're doing this together w DAMIT so at the end of the road there's one data catalog for AI datasets across Meta

**Commented [13]:** why have we not touched on llm so far?

**Commented [14]:** why is there no info on scale? How are we supporting nextgen models?

| | | |
|---|---|---|
| Q4-23 | • All technical **blockers removed** for authoring and moving **LLM** pipelines to Infra<br>• **Migrate** existing **Dedup & UII** redaction algorithms to reusable **Dataswarm** Operators<br>• Data **Catalog** available for registration of **all datasets**<br>• **Video** and **audio** dataset **visualization** (single view for videos / audio and associated metadata)<br>• Dataset **comparison** tool<br>• ***Canonical**, fully annotated, media datasets (dim_all_photos / dim_all_videos) | • **PySpark** available for production usage<br>• **Dataswarm** Operators for data deduplication and UII redaction<br>• AI Data Catalog extensions for dataset **visualization** and **comparison**<br>• ***Daily pipeline** for **annotating** new photos and videos, leveraging DPS / Batch Inference / MAS |

## Discussion & Next steps

1. **Have we prioritized the right investments? Are there any other pain points we should look into immediately and longer term?**
   a. We're looking to further develop the roadmap for H2'23 and beyond with you, and get back with the specific efforts and respective ETA to best address your needs.

2. **Should Infra teams (AI Infra, Data Infra, Core Data, Storage) and GenAI invest in:**
   a. **Canonical photos and videos datasets**, to be used as training data candidates and search & retrieval data?
      i. The GAID team has mitigated and annotated ~10% of the Instagram photos and videos. Infra teams have partnered with GAID on running the inference pipelines for annotations. These require significant GPU capacity, which is not available short term. As model evaluation is proceeding and shortcomings are identified, more data is annotated on demand.
      ii. ~~[Is there strategic value in defining dim_all_photos / dim_all_videos datasets, with mitigations and augmentations needed for GenAI, to enable faster data selection and iteration in the future?]~~
      iii. We may be able to leverage CPUs instead of GPUs for most annotations (not BLIP). Do we see value in exploring accelerating photo & media selection for training?
   b. **Very large scale and real-time embedding store** for search & retrieval?
      i. We do not have a unified large scale vector database offering like Pinecone. Such databases have been growing in popularity across the industry, for both Ranking and Recommendations candidate generators and Generative AI inference-time search & retrieval.
      ii. GenAI teams have already built IG-KNN, Text-KNN on top of existing technologies. Do you see any challenges with existing Infra solutions? Do you see a need for a larger strategic investment in this area, like a consolidated solution across all documents?

3. **Help needed to ensure good tradeoffs between immediate and long term focus:**
   a. Continued feedback and guidance, as we build and adopt new solutions (i.e. Data Preparation Service, AI Data Catalog, PySpark, new Dataswarm Operators).

**Commented [15]:** why is there no info on scale? How are we supporting nextgen models?

**Commented [16]:** This should be agnostic of photos or videos, there is a serious privacy/mitigation issue which current framework/tech stack just doesn't solve in a scalable fashion.

**Commented [17]:** Here we were referring to our large database of photos and videos without signals / annotations, and whether annotating a large fraction of them would speed up certain selection operations and possibly open up to higher quality datasets (where we indeed share the concern Roshan brought up, that we do not understand the ROI, so we have a chicken-and-egg problem).

Could you share more on the mitigation concern y ... [3]

**Commented [18]:** We had discussed this in another review where this investment question can only be answered by understanding how scaling up media data (for pre-training) impacts foundation model quality ... [4]

**Commented [19]:** cc [redacted]@meta.com [redacted]@fb.com

**Commented [20]:** I don't think we will need to have one giant dimension table with all images / videos. We would rather need a Media data mart, similar to the one we are planning to build on Languages side (cc: ... [5]

**Commented [21]:** We have two questions:
- do we see value in maintaining up-to-date tables of photos / videos including common GenAI annotations (such as entities / watermark / aesthetics / maybe ... [6]

**Commented [22]:** We have signals table and curated dataset table. Our intuition is for the curated dataset, quality is likely more important than quantity (to be validated by future experiments). ... [7]

**Commented [23]:** I think now it's worth directing some help to the video data side, where the data scale still needs to grow, and basic signal like XrayVideo is being throttled. ... [8]

**Commented [24]:** If the focus here on "online vector database", we do have LaserKNN (built on FAISS) and LCC (also wrapped in Laser API) - both of which do a fantastic job in-house. ... [9]

**Commented [25]:** Right now the only "online retrieval" use-case where such a vector database would be useful is for safety - but that too is readily available via other services like MMS (Media Matching Servic ... [10]

**Commented [26]:** Thanks for the details! Laser kNN updates its centroids on a periodic basis (not realtime, more like daily) using a DiGraph pipeline. This is what originated the question. ... [11]

**Commented [27]:** We don't need real-time. Larger-scale could be useful. Our laser capacity is now 10-15B, that limits the amount of data we can look up from. If we ever need to use text KNN for LLM d ... [12]

b. Visibility into GenAI timelines and evolving priorities (i.e. future workload understanding, including types of multimodality and data usage patterns).

**Page 3: [1] Commented [6]** **Delia David** **8/9/2023 5:26:00 AM**

This is a new system more tuned for AI datasets vs the existing Metastore. Functionality built here will be around multi-table datasets, versioned access and metadata support for modality tracking / mitigations / annotations.

Given the context above, this system will be integrated in the data preparation and training workflows in H2. Please find a few more below (also please note that we review all these in a sprint based fashion with GenAI - and we have not reached this milestone yet, so please take this as tentative integration):

- Data Preparation: as data is ingested into our data warehouse, mitigated and transformed, metadata on these operations is automatically. tracked when using Infra-provided Dataswarm Operators (which we will design and build together). For non-standard Operators, we will have APIs to register the same pieces of metadata.

- Training workflows: imagine using a dataset backed by multiple tables. Say we call it X. Training job owners would go to Mast and specify dataset X in their job config. Mast would resolve X into underlying tables by calling this new service. This way, customers do not need to remember and type 4 table names when using a dataset with 4 tables.

- Debugging in Presto / Spark: we have not planned this yet, but we might want to add an integration so that one can go to DaiQuery and use a dataset name for querying vs the 4 underlying tables in the example above.

**Page 3: [2] Commented [8]** **Chaya Nayak** **8/9/2023 3:46:00 PM**

**Redacted - Privilege**

1 total reaction
Parth Parekh reacted with at 2023-08-09 10:56 AM

**Page 7: [3] Commented [17]** **Delia David** **8/9/2023 5:35:00 AM**

Here we were referring to our large database of photos and videos without signals / annotations, and whether annotating a large fraction of them would speed up certain selection operations and possibly open up to higher quality datasets (where we indeed share the concern Roshan brought up, that we do not understand the ROI, so we have a chicken-and-egg problem).

Could you share more on the mitigation concern you are referring to and how this is related? My understanding is that at least the pipelines from GAID (Vladan, Guan and team) are properly mitigated for privacy.

**Page 7: [4] Commented [18]** **Roshan Sumbaly** **8/9/2023 3:53:00 AM**

We had discussed this in another review where this investment question can only be answered by understanding how scaling up media data (for pre-training) impacts foundation model quality.

Unfortunately, unlike a thorough ablation done on the language side on this topic [1], similar study on media side is missing - which makes this a difficult question to answer without some validation of the usefulness of even larger dataset.

[1] - https://arxiv.org/abs/2203.15556
2 total reactions
Delia David reacted with □ at 2023-08-08 22:17 PM
Parth Parekh reacted with □ at 2023-08-08 21:47 PM

**Page 7: [5] Commented [20]** **Parth Parekh** **8/9/2023 4:48:00 AM**

I don't think we will need to have one giant dimension table with all images / videos. We would rather need a Media data mart, similar to the one we are planning to build on Languages side (cc: @[redacted]@meta.com) with all the metadata stored in one mart for faster analytics.

**Page 7: [6] Commented [21]** **Delia David** **8/9/2023 5:09:00 AM**

We have two questions:
- do we see value in maintaining up-to-date tables of photos / videos including common GenAI annotations (such as entities / watermark / aesthetics / maybe CLIP) -- for easy selection of data as datasets need to be extended

- how we organize data, one or multiple tables. While I called it one table, I do expect that even from a privacy point of view we might need a few tables, organized around our products. And we might only be able to use certain subsets for certain models.

The Chinchilla paper talks about scale laws and how smaller models can do as good as larger models, provided they have high quality data. The question here is more about the exact data and its quality vs the quantity of data. One concrete example being needing to extend a dataset with new data as we identify certain prompts (on specific classes of entities) do not work well and we need better quality data or data which was previously missing or under sampled. Concretely, if we need more high resolution mountain images, do we already find it easy to identify such photos today (in the ~10% IG annotated set, extending to FB this half) or would we benefit from a fully annotated dataset, assuming we figure out a capacity story for running the annotations?

And if I read Roshan's comment well, we have no evidence internally and in research we are aware of, that we need the above. The current datasets enable the performance we need, I assume on par or better with other similar products like Dall-e/Midjourney/Runway/etc.
1 total reaction
Parth Parekh reacted with  at 2023-08-09 10:57 AM

**Page 7: [7] Commented [22]** **Guan Pang** **8/9/2023 3:20:00 PM**

We have signals table and curated dataset table. Our intuition is for the curated dataset, quality is likely more important than quantity (to be validated by future experiments).

But here we're talking about the signals table, which could benefit from bigger scale that allows us more chance to select a higher quality curated dataset of the same size.

That said, this also depends on how fast we can improve our data curation method to be able to benefit from larger signals table (since naively increasing thresholds will usually result in more bias).
1 total reaction
Delia David reacted with  at 2023-08-09 08:37 AM

**Page 7: [8] Commented [23]** **Guan Pang** **8/9/2023 3:25:00 PM**

I think now it's worth directing some help to the video data side, where the data scale still needs

to grow, and basic signal like XrayVideo is being throttled.
1 total reaction
Delia David reacted with □ at 2023-08-09 08:37 AM

**Page 7: [9] Commented [24]** **Roshan Sumbaly** **8/9/2023 3:44:00 AM**

If the focus here on "online vector database", we do have LaserKNN (built on FAISS) and LCC (also wrapped in Laser API) - both of which do a fantastic job in-house.

Of course not to be confused with what we're doing for IG-KNN/Text-KNN where the goal isn't "real-time retrieval" - more for near-real-time quick look-ups

**Page 7: [10] Commented [25]** **Roshan Sumbaly** **8/9/2023 3:47:00 AM**

Right now the only "online retrieval" use-case where such a vector database would be useful is for safety - but that too is readily available via other services like MMS (Media Matching Service, which under the hood use LaserKNN/LCC that I mentioned above).

**Page 7: [11] Commented [26]** **Delia David** **8/9/2023 5:12:00 AM**

Thanks for the details! Laser kNN updates its centroids on a periodic basis (not realtime, more like daily) using a DiGraph pipeline. This is what originated the question.

It does look like this is already great for your use cases and we do not need more realtime updates at the moment.

**Page 7: [12] Commented [27]** **Guan Pang** **8/9/2023 3:28:00 PM**

We don't need real-time. Larger-scale could be useful. Our laser capacity is now 10-15B, that limits the amount of data we can look up from. If we ever need to use text KNN for LLM data I imagine we need much larger capacity too.

_.docx

Main document changes and comments

**Page 1: Inserted** **Jeff Marcus** **9/13/2023 5:04:00 PM**

feel

**Page 2: Commented [1]** **Manohar Paluri** **8/9/2023 3:44:00 PM**

can i get some steer and can we touch on how our media and language folks on how this has helped us so far?

**Page 2: Commented [2]** **Manohar Paluri** **8/9/2023 3:43:00 PM**

this reads like it is not done and we need to?

**Page 2: Commented [3]** **Manohar Paluri** **8/9/2023 3:44:00 PM**

where are we on language?

**Page 3: Commented [4]** **Manohar Paluri** **8/9/2023 3:45:00 PM**

i think we should really optimize for:
1. scale
2. access on TC
3. streamlining pxfn decisions because of the data catalog effort.

**Page 3: Commented [5]** **Parth Parekh** **8/9/2023 4:50:00 AM**

Do we have more details on this in terms of data flow and how does it fit with existing workflows?

**Page 3: Commented [6]** **Delia David** **8/9/2023 5:26:00 AM**

This is a new system more tuned for AI datasets vs the existing Metastore. Functionality built here will be around multi-table datasets, versioned access and metadata support for modality tracking / mitigations / annotations.

Given the context above, this system will be integrated in the data preparation and training workflows in H2. Please find a few more below (also please note that we review all these in a sprint based fashion with GenAI - and we have not reached this milestone yet, so please take this as tentative integration):

- Data Preparation: as data is ingested into our data warehouse, mitigated and transformed, metadata on these operations is automatically. tracked when using Infra-provided Dataswarm Operators (which we will design and build together). For non-standard Operators, we will have APIs to register the same pieces of metadata.

- Training workflows: imagine using a dataset backed by multiple tables. Say we call it X. Training job owners would go to Mast and specify dataset X in their job config. Mast would resolve X into underlying tables by calling this new service. This way, customers do not need to remember and type 4 table names when using a dataset with 4 tables.

- Debugging in Presto / Spark: we have not planned this yet, but we might want to add an integration so that one can go to DaiQuery and use a dataset name for querying vs the 4

underlying tables in the example above.

**Page 3: Commented [7]** **Chaya Nayak** **8/9/2023 3:45:00 PM**

Curious if and how this interfaces with idata?

**Page 3: Commented [8]** **Chaya Nayak** **8/9/2023 3:46:00 PM**

**Redacted - Privilege**

1 total reaction
Parth Parekh reacted with at 2023-08-09 10:56 AM

**Page 3: Commented [9]** **David Levin** **8/10/2023 8:46:00 AM**

@[redacted]@meta.com - yes, this is possible and where we'd like to go. What would be useful is to understand the process in more details, worth a followup.

re iData - yes, this interfaces w iData and is actually built by few members of the iData team.

**Page 6: Commented [10]** **Jeet Shah** **8/9/2023 5:15:00 AM**

How will this be different from DAMIT?

Furthmore, shouldn't DE/DS own this?

**Page 6: Commented [11]** **Delia David** **8/9/2023 5:30:00 AM**

Part of the visual capabilities and insights exist in DAMIT. DAMIT is working with AI Infra and Data Infra on building together and consolidating onto the new Data Catalog solution, which will offer higher scale and more capabilities over time.

I assume that a lot of data insights are being built by DE/DS today. We want to have a discussion on whether we see value in computing some of these by default for certain classes of datasets. This would only make sense if the same insights apply across a large number of datasets.

**Page 6: Commented [12]** **David Levin** **8/9/2023 10:54:00 AM**

We also look to expand the existing Media data understanding capabilities in DAMIT to other modalities including multimodal and LLM. We're doing this together w DAMIT so at the end of the road there's one data catalog for AI datasets across Meta

**Page 6: Commented [13]** **Manohar Paluri** **8/9/2023 3:46:00 PM**

why have we not touched on llm so far?

**Page 6: Commented [14]** **Manohar Paluri** **8/9/2023 3:47:00 PM**

why is there no info on scale?
How are we supporting nextgen models?

**Page 7: Commented [15]** **Manohar Paluri** **8/9/2023 3:47:00 PM**

why is there no info on scale?
How are we supporting nextgen models?

**Page 7: Commented [16]** **Jeet Shah** **8/9/2023 5:14:00 AM**

This should be agnostic of photos or videos, there is a serious privacy/mitigation issue which current framework/tech stack just doesn't solve in a scalable fashion.

**Page 7: Commented [17]** **Delia David** **8/9/2023 5:35:00 AM**

Here we were referring to our large database of photos and videos without signals / annotations, and whether annotating a large fraction of them would speed up certain selection operations and possibly open up to higher quality datasets (where we indeed share the concern Roshan brought up, that we do not understand the ROI, so we have a chicken-and-egg problem).

Could you share more on the mitigation concern you are referring to and how this is related? My understanding is that at least the pipelines from GAID (Vladan, Guan and team) are properly mitigated for privacy.

**Page 7: Commented [18]** **Roshan Sumbaly** **8/9/2023 3:53:00 AM**

We had discussed this in another review where this investment question can only be answered by understanding how scaling up media data (for pre-training) impacts foundation model quality.

Unfortunately, unlike a thorough ablation done on the language side on this topic [1], similar study on media side is missing - which makes this a difficult question to answer without some validation of the usefulness of even larger dataset.

[1] - https://arxiv.org/abs/2203.15556
2 total reactions
Delia David reacted with at 2023-08-08 22:17 PM
Parth Parekh reacted with at 2023-08-08 21:47 PM

**Page 7: Commented [19]** **Roshan Sumbaly** **8/9/2023 3:54:00 AM**

cc @[redacted]@meta.com @[redacted]@fb.com

**Page 7: Commented [20]** **Parth Parekh** **8/9/2023 4:48:00 AM**

I don't think we will need to have one giant dimension table with all images / videos. We would rather need a Media data mart, similar to the one we are planning to build on Languages side (cc: @[redacted]@meta.com) with all the metadata stored in one mart for faster analytics.

**Page 7: Commented [21]** **Delia David** **8/9/2023 5:09:00 AM**

We have two questions:
- do we see value in maintaining up-to-date tables of photos / videos including common GenAI annotations (such as entities / watermark / aesthetics / maybe CLIP) -- for easy selection of data as datasets need to be extended

- how we organize data, one or multiple tables. While I called it one table, I do expect that even from a privacy point of view we might need a few tables, organized around our products. And we might only be able to use certain subsets for certain models.

The Chinchilla paper talks about scale laws and how smaller models can do as good as larger models, provided they have high quality data. The question here is more about the exact data and its quality vs the quantity of data. One concrete example being needing to extend a dataset with new data as we identify certain prompts (on specific classes of entities) do not work well and we need better quality data or data which was previously missing or under sampled. Concretely, if we need more high resolution mountain images, do we already find it easy to identify such photos today (in the ~10% IG annotated set, extending to FB this half) or would we benefit from a fully annotated dataset, assuming we figure out a capacity story for running the annotations?

And if I read Roshan's comment well, we have no evidence internally and in research we are aware of, that we need the above. The current datasets enable the performance we need, I assume on par or better with other similar products like Dall-e/Midjourney/Runway/etc.
1 total reaction
Parth Parekh reacted with □ at 2023-08-09 10:57 AM

**Page 7: Commented [22]** **Guan Pang** **8/9/2023 3:20:00 PM**

We have signals table and curated dataset table. Our intuition is for the curated dataset, quality is likely more important than quantity (to be validated by future experiments).

But here we're talking about the signals table, which could benefit from bigger scale that allows us more chance to select a higher quality curated dataset of the same size.

That said, this also depends on how fast we can improve our data curation method to be able to benefit from larger signals table (since naively increasing thresholds will usually result in more bias).
1 total reaction
Delia David reacted with at 2023-08-09 08:37 AM

**Page 7: Commented [23]** **Guan Pang** **8/9/2023 3:25:00 PM**

I think now it's worth directing some help to the video data side, where the data scale still needs to grow, and basic signal like XrayVideo is being throttled.
1 total reaction
Delia David reacted with at 2023-08-09 08:37 AM

**Page 7: Commented [24]** **Roshan Sumbaly** **8/9/2023 3:44:00 AM**

If the focus here on "online vector database", we do have LaserKNN (built on FAISS) and LCC (also wrapped in Laser API) - both of which do a fantastic job in-house.

Of course not to be confused with what we're doing for IG-KNN/Text-KNN where the goal isn't "real-time retrieval" - more for near-real-time quick look-ups

**Page 7: Commented [25]** **Roshan Sumbaly** **8/9/2023 3:47:00 AM**

Right now the only "online retrieval" use-case where such a vector database would be useful is for safety - but that too is readily available via other services like MMS (Media Matching Service, which under the hood use LaserKNN/LCC that I mentioned above).

**Page 7: Commented [26]** **Delia David** **8/9/2023 5:12:00 AM**

Thanks for the details! Laser kNN updates its centroids on a periodic basis (not realtime, more like daily) using a DiGraph pipeline. This is what originated the question.

It does look like this is already great for your use cases and we do not need more realtime updates at the moment.

**Page 7: Commented [27]** **Guan Pang** **8/9/2023 3:28:00 PM**

We don't need real-time. Larger-scale could be useful. Our laser capacity is now 10-15B, that limits the amount of data we can look up from. If we ever need to use text KNN for LLM data I imagine we need much larger capacity too.

Header and footer changes

Text Box changes

Header and footer text box changes

Footnote changes

Endnote changes